Exhibit A

Introduction to Protein Structure

Carl Branden
Uppsala Biomedical Center
Swedish University of Agricultural Sciences
Uppsala, Sweden

John Tooze
European Molecular Biology Organization
Heidelberg, Germany

Garland Publishing, Inc.
New York and London, 1991

Protein molecules are organized in a structural hierarchy

The Danish biochemist Kai Linderstrøm-Lang coined the terms primary, secondary, and tertiary structure to emphasize the structural hierarchy in proteins (see Figure 1.1). **Primary structure** is the linear amino acid sequence; in other words, the arrangement of amino acids along a polypeptide chain. When two different proteins have significant similarities in their primary structures, they are said to be homologous to each other. Since the corresponding DNA sequences also are significantly similar, it is generally assumed that they are evolutionarily related; they have evolved from a common ancestoral gene.

Secondary structure occurs mainly as α helices and β strands. The formation of secondary structure in a local region of the polypeptide chain is to some extent determined by the primary structure. Certain amino acid sequences favor either α helices or β strands; others favor formation of loop regions. Secondary structure elements usually arrange themselves in simple motifs, as described above. Motifs are formed by packing side chains from adjacent α helices or β strands close to each other.

Several motifs usually combine to form compact globular structures, which are called **domains**. In this book we will use **tertiary structure** as a common term both for the way motifs are arranged into domain structures and for the way a single polypeptide chain folds into one or several domains. Domains with homologous amino acid sequences in different proteins almost invariably have similar tertiary structures.

During the last decades the central dogma of protein folding has stated that the primary structure determines the tertiary structure; in other words, the final specific tertiary structure forms spontaneously from a polypeptide chain with a unique amino acid sequence. Recently, however, a set of proteins, called **chaperones**, have been identified that are required for formation of proper tertiary structure of many different proteins. They seem to act as catalysts by increasing the rates of the final steps in the folding process, and thus in principle they do not violate the central dogma of folding.

Many protein molecules have only one chain; these are monomeric proteins. But a fairly large number have several identical polypeptide chains that associate into a multimeric molecule with a specific **quaternary structure**. These subunits can function either independently of each other or cooperatively so that the function of one subunit is dependent on the functional state of other subunits. Other protein molecules are assembled from several different subunits with different functions, for example, RNA polymerase from *E. coli*, which contains five different polypeptide chains.

Large polypeptide chains fold into several domains

The fundamental unit of tertiary structure is the **domain**. A domain is defined as a polypeptide chain or a part of a polypeptide chain that can independently fold into a stable tertiary structure. Domains are also units of function. Often, the different domains of a protein are associated with different functions. For example, in the lambda repressor protein, discussed in Chapter 7, there is one domain at the N terminus of the polypeptide chain that binds DNA and a second C-terminal domain that holds two polypeptide chains together into a dimeric repressor molecule.

Proteins may comprise a single domain or as many as several dozen domains (Figure 2.19). There is no fundamental structural distinction between a domain and a subunit; there are many known examples where several biological functions that are carried out by separate polypeptide chains in one species are performed by domains of a single protein in another species. For example, synthesis of fatty acids requires catalysis of seven different chemical reactions. In plant chloroplasts these reactions are catalyzed by seven different proteins, whereas in mammals they are performed by one polypeptide chain arranged in seven domains. Such differences thus reflect the organization of the genome rather than the dictates of structure.

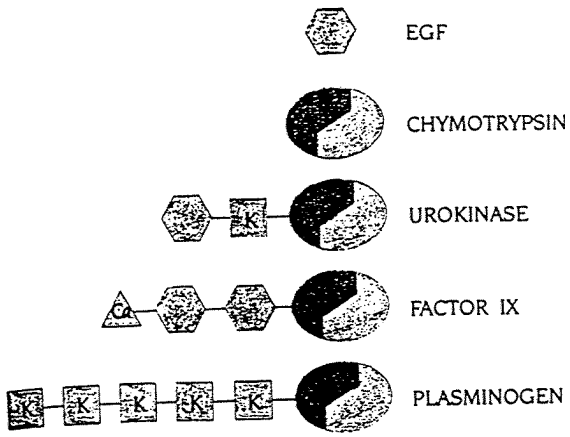

Figure 2.19 Organization of polypeptide chains into domains. Small protein molecules like the epidermal growth factor, EGF, comprise only one domain. Others like the serine proteinase chymotrypsin are arranged in two domains that are both required to form a functional unit (Chapter 15). Many of the proteins that are involved in blood coagulation and fibrinolysis, such as urokinase, factor IX, and plasminogen have long polypeptide chains that comprise different combinations of domains homologous to EGF and serine proteinases and, in addition, calcium-binding domains and Kringle domains.

Domains that are homologous to the epidermal growth factor, EGF, which is a small polypeptide chain of 53 amino acids;

Serine proteinase domains that are homologous to chymotrypsin, which has about 245 amino acids arranged in two domains;

Kringle domains that have a characteristic pattern of three internal disulphide bridges within a region of about 85 amino acid residues;

Calcium-binding domain (see Figure 2.13).

Domains are built from structural motifs

Domains are formed by different combinations of secondary structure elements and motifs. The α helices and β strands of the motifs are adjacent to each other in the three-dimensional structure and connected by loop regions. Sequentially adjacent motifs, in other words, motifs that are formed from consecutive regions of the primary structure of a polypeptide chain, are usually close together in the three-dimensional structure (Figure 2.20). Thus to a first approximation a polypeptide chain can be sequentially arranged in a number of these simple motifs. The number of such combinations found in proteins is limited, and some combinations seem to be structurally favored. Thus similar domain structures frequently occur in different proteins with different functions and with completely different amino acid sequences.

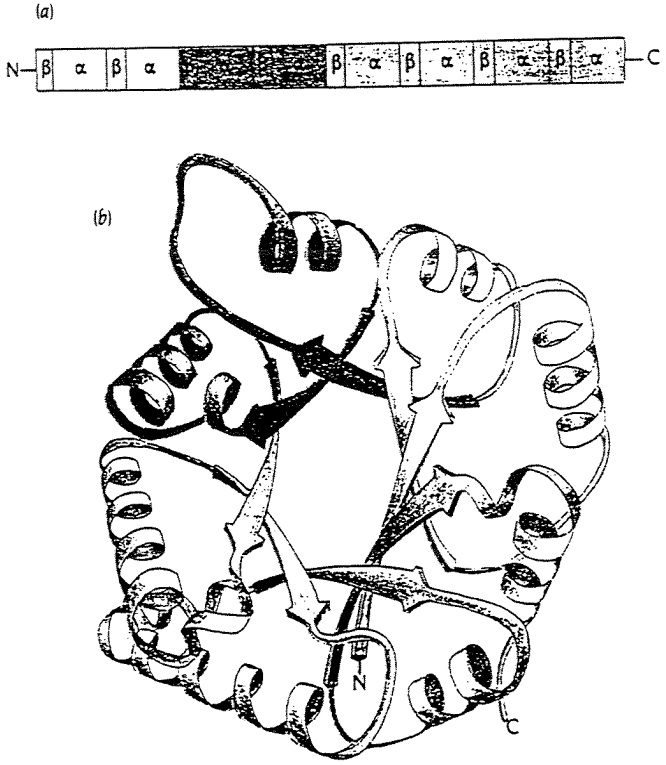

Figure 2.20 Motifs that are adjacent in the amino acid sequence are also usually adjacent in the three-dimensional structure. Triose phosphate isomerase (Figure 2.10) is built up from four β-α-β-α motifs that are consecutive both in the amino acid sequence (a) and in the three-dimensional structure (b).

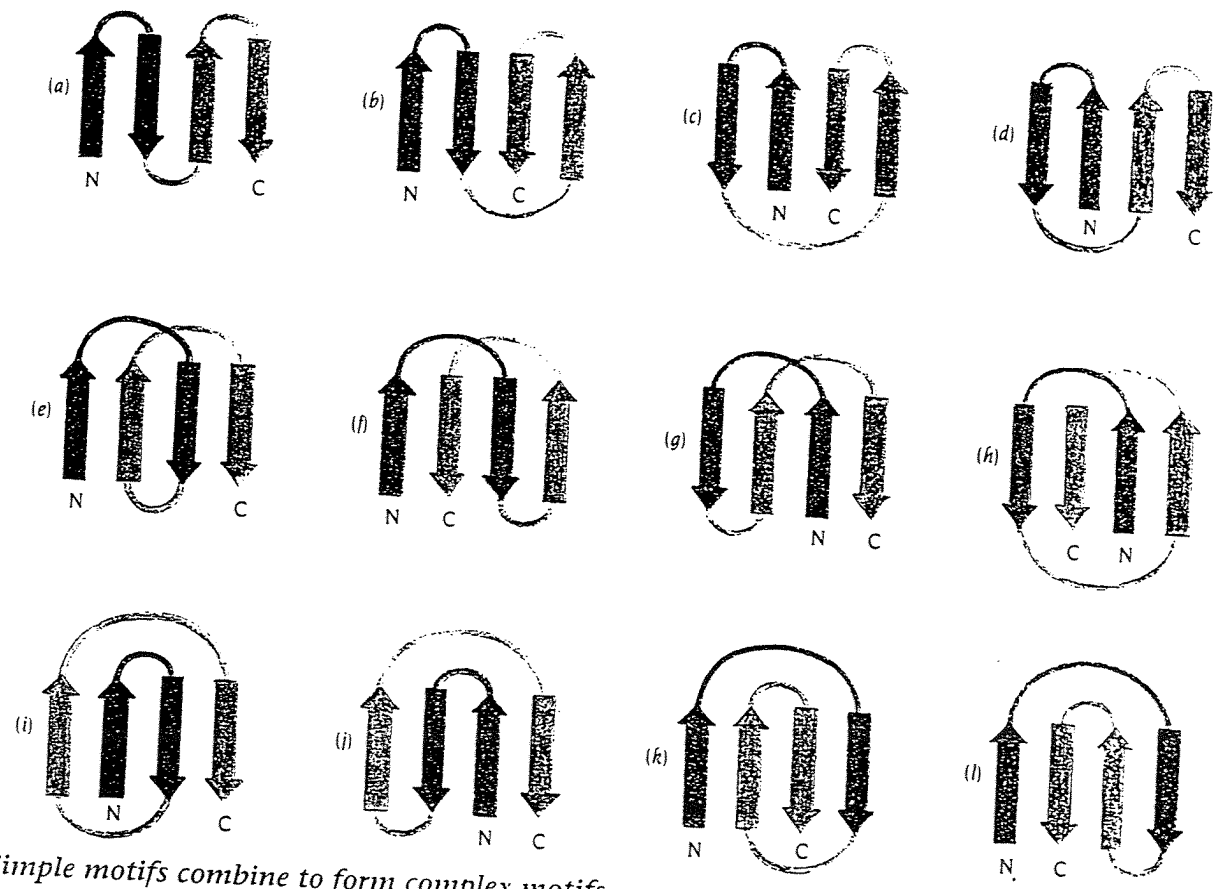

Figure 2.21 Two sequentially adjacent hairpin motifs can be arranged in 12 different ways into a β sheet of four strands. (a) Topology diagram for the arrangement when sequentially adjacent β strands are adjacent in the structure. (b–l) are other possible arrangements, and of these j and l are the Greek key motifs. The topology diagram in j is the mirror image of that in Figure 2.15a, and the diagram in l is the same as that in Figure 2.16 but with the chain direction reversed. These four diagrams represent the Greek key motif.

Simple motifs combine to form complex motifs

Figure 2.21 illustrates all the possible ways in which two adjacent β hairpin motifs, each consisting of two antiparallel β strands connected by a loop region, can be combined to make a more complex motif.

The simplest arrangement is that shown in Figure 2.21a. This occurs in many proteins and is illustrated by the structure of one domain of aspartate transcarbamylase (Figure 2.11a).

There are 11 additional ways to build up a β sheet of four strands from these two units. In a random sample of 10 β structures the arrangement in Figure 2.21a occurs 11 times out of a total number of 27 structural fragments of four adjacent β strands. In addition, arrangements 2.21j and 2.21l occur 6 times each. These are the Greek key motifs, previously discussed. Most of the theoretically possible arrangements do not occur within this sample of structures and some like 2.21b and 2.21d occur only once or twice.

Protein structures can be divided into three main classes

Based on simple considerations of connected motifs, Michael Levitt and Cyrus Chothia of the M.R.C. Laboratory of Molecular Biology in Cambridge derived a taxonomy of protein structures, and they could show that combinations of these motifs build up the core of most domain structures and also form the basis for a classification into three main groups: α domains, β domains, and α/β domains. In α structures the core is built up exclusively from α helices (Figure 2.9), and in β structures the core comprises antiparallel β sheets, usually two β sheets packed against each other (Figure 2.11c). The α/β structures are made from combinations of β-α-β motifs that form a predominantly parallel β sheet surrounded by α helices (Figures 2.10b and 2.11b).

Some proteins are built up from a combination of discrete α and β motifs and usually form one small antiparallel β sheet in one part of the domain packed against a number of α helices (Figure 2.15). These structures can be considered

to belong to a small fourth group. In addition to these groups, there are a number of small proteins that are rich in disulfide bonds or metals and form a special group. The structures of these proteins seem to be strongly influenced by the presence of these metals or disulfides and often look like distorted versions of more regular proteins.

The domains of the known protein structures are classified according to Levitt and Chothia's scheme in this book. The three main classes α, β, and α/β will be examined in more detail in Chapters 3–5.

Conclusion

The interiors of protein molecules contain mainly hydrophobic side chains. The main chain in the interior is arranged in secondary stuctures to neutralize its polar atoms through hydrogen bonds. There are two main types of secondary structure, α helices and β sheets. β sheets can have their strands parallel, antiparallel, or mixed.

Protein structures are built up by combinations of secondary structural elements, α helices, and β strands. These form the core regions—the interior of the molecule—and they are connected by loop regions at the surface. Schematic and simple topological diagrams where these secondary structure elements are highlighted are very useful and are frequently used. α helices or β strands that are adjacent in the amino acid sequence are also usually adjacent in the three-dimensional structure. Certain combinations are especially frequent and are called motifs, for example, the helix-loop-helix motif and the hairpin motif. Two helix-loop-helix motifs, each with its own specific geometry and amino acid sequence requirements, are used in many different proteins, one motif for DNA binding and one for calcium binding.

The β-α-β motif, which consists of two parallel β strands joined by an α helix, occurs in almost all structures that have a parallel β sheet. Four antiparallel β strands, arranged in a specific way, the Greek key motif, are frequently found in structures with antiparallel β sheets.

Polypeptide chains are folded into one or several discrete units, domains, which are the fundamental functional and three-dimensional structural units. The cores of domains are built up from combinations of small motifs of secondary structure, such as α-loop-α, β-loop-β or β-α-β motifs. Domains are classified into three main structural groups: α structures, where the core is built up exclusively from α helices; β structures, which comprise antiparallel β sheets; and α/β structures, where combinations of β-α-β motifs form a predominantly parallel β sheet surrounded by α helices.